DOCUMENT RESUME

ED 300 454 TM 012 504

AUTHOR Kingston, Neal M.
TITLE The Effect of Scoring Instructions and Item Type on Lower Asymptotes in the Three-Parameter Logistic Model.
PUB DATE Mar 85
NOTE 16p.; An earlier version of this paper presented at the Meeting of the National Council on Measurement in Education (San Francisco, CA, March 31-April 4, 1985).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *College Entrance Examinations; Guessing (Tests); Higher Education; *Item Analysis; *Scoring; Testing Problems; *Verbal Tests
IDENTIFIERS Antonyms; Asymptotic Distributions; *Graduate Record Examinations; *Rights and Formula Scoring; Test Instructions; Three Parameter Model

ABSTRACT

This research investigated the effect on estimated lower asymptotes of the instructions to Graduate Record Examination (GRE) examinees about how the test would be scored. This effect was assessed for four different verbal item types (analogies, antonyms, sentence completion, and reading comprehension) using a two-way, unweighted means analysis of variance on estimated c-parameters--the probability that an extremely low ability examinee will get the item correct (chance probability). The LOGIST program was used to estimate item and person parameters for the GRE administered under formula-scoring and right-scoring instructions. Sample sizes ranged from 2,900 to 4,800. It was thought that the probability of a correct response by low ability examinees might systematically differ under rights- and formula-score instructions; this did not occur. It was also thought that the mean c-parameters might vary by item type; this did occur. Most notably, despite much evidence that estimated c's are typically considerably less than the reciprocal of the number of choices for the item, one item type--antonyms--was identified for which this is not so. It is recommended that these findings be taken into account when estimating item parameters. Four data tables conclude the document. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300454

THE EFFECT OF SCORING INSTRUCTIONS AND ITEM TYPE ON LOWER ASYMPTOTES IN THE THREE-PARAMETER LOGISTIC MODEL

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Neal M. Kingston

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Neal M. Kingston
Educational Testing Service

Address all correspondence to:
Neal Kingston, 20-P
Educational Testing Service
Princeton, NJ 08541

TM012 504

ABSTRACT

This research investigated the effect on estimated lower asymptotes of the instructions to examinees about how the test would be scored. This was assessed for four different verbal item types using a two-way, unweighted means analysis of variance. It was thought that the probability of a correct response by low ability examinees might systematically differ under rights-and formula-score instructions. This did not occur. It was also thought the mean c-parameters might vary by item type. This did occur. Most notably, despite much evidence that estimated c's are typically considerably less than the reciprocal of the number of choices for the item, an item type, antonyms, was identified for which this is not so. It is recommended that these findings be taken into account when estimating item parameters.

INTRODUCTION

When an examinee taking an ability or achievement test is faced with an item for which he or she is not sure as to the correct answer, a complex decision making process might occur. Assuming that the examinee wants to obtain as high a score as possible, given the scoring instructions for the test, when faced with an item for which the correct response is unclear, the examinee can determine and follow some strategy to maximize her or his score. This strategy will be affected by partial information and misinformation that the examinee may have. Finally, examinees typically are not purely rational decision theorists. Various personality traits affect an examinee's behavior in the face of uncertainty.

PURPOSE

In October 1981 the GRE General Test (called the Aptitude Test until October 1982) switched from using formula-scoring instructions to right-scoring instructions. In order to explore the effect of scoring instructions on item responding behavior, this paper addresses several questions: Did this change affect the probability of responding correctly to an item for very low ability examinees? Also, are there any consistent differences in the probability of a correct response for very low ability examinees for the four GRE verbal measure item types. Finally, can any hypotheses about examinee guessing behavior be generated from observed differences in probabilities of correct responses of low ability examinees.

THE THREE-PARAMETER ITEM RESPONSE MODEL

The three-parameter logistic model (Birnbaum, 1968) assumes that for an examinee of given ability, theta, three statistical aspects of the an item determine the probability that the examinee will respond correctly: a, the discriminating power of the item; b, the difficulty of the item; and c, the lower asymptote of the item response function. The c-parameter represents the probability that an extremely low ability examinee (theta approaching negative infinity) will get the item correct. The c-parameter has been referred to as a guessing parameter, but since for most multiple-choice items its value is less than the chance probability of a correct response (that is, 1/A, where A is the number of response options), which is what would occur with random guessing, it is more frequently referred to as a pseudo-guessing parameter, or simply as a lower asymptote parameter.

TEST EDITIONS AND SAMPLES

The GRE General Test verbal measure consists of four item types: analogies, antonyms, sentence completion, and reading comprehension. Description and examples of these item types can be found in any edition of the GRE Information Bulletin (e.g., ETS, 1984). The verbal measure, as did the other General Test measures, underwent several changes as of October 1981. Foremost, the scoring instructions changed from formula (rights minus one-quarter wrongs) to number-right. Thus, it was more clearly in the examinees' interests to guess when they were unsure of the answer to an item for the post-October 1981 verbal measure.

Table 1 presents for each test edition, the administration date, number of analogy, antonym, sentence completion, and reading comprehension items, the overall difficulty of the verbal measure (mean equated delta), the mean scaled score of the sample, and the sample size. All items have five response options.

Insert Table 1 About Here

DATA ANALYSIS

The program LOGIST was used to estimate item and person parameters based on the three-parameter logistic model for four editions of the GRE General Test administered between December 1979 and June 1980 under formula-scoring instructions, and for six editions administered between October 1981 and April 1983 under right-scoring instructions. Sample sizes ranged from about 2,900 to 4,800. This paper presents comparisons of estimated c-parameters for the four GRE verbal item types: analogies, antonyms, sentence completion, and reading comprehension, administered under formula- and right-scoring instructions.

It should be noted that there were two important differences in the estimation procedures for the c-parameters that might have influenced the results of this study. Both relate to the procedures LOGIST uses to estimate the c-parameter for items that have insufficient data at the lower asymptote. LOGIST allows the user to decide how much data is "enough" for estimating c. At the time the parameters were estimated for the formula-scored tests, the

author was conservative and there were many items for which it was decided there was insufficient data to estimate a unique c. After obtaining more experience with both LOGIST and GRE data, when LOGIST was used to estimate parameters for the right-scored tests, a less conservative approach was used and a unique c was estimated for a considerably larger proportion of the items. This is reflected in the difference between the standard deviations for the two scoring instruction conditions (see Table 2). Perhaps more critically, the procedure for estimating the "common c" for those items for which there were not sufficient data differed. The version of LOGIST used to estimate parameters for the formula-scored tests used the median of the c-parameter estimates of those items for which there were unique estimates. The more recent version of LOGIST used to estimate parameters for the right-scored tests estimated the common c using modified maximum likelihood methods based on combined data for all such items (Wingersky, 1983).

To determine the effect of the change from formula-scoring to right-scoring instructions for the four GRE verbal item types, a two-way, unweighted means analysis of variance was performed on the estimated c-parameters (Winer, 1971, chapter 5.22).

RESULTS

Table 2 presents the standard deviations for each cell in the ANOVA. Although the standard deviations show clearly that the assumption of homogeneous variances is violated, it has been shown that ANOVA is robust to violations much more severe than this (Box, 1954).

Table 3 presents the mean of the estimated c-parameters for each item type and scoring instruction condition. Table 4 presents the analysis of variance. The differences attributable to scoring instructions are very small and are not statistically significant at any commonly accepted level. The differences among the four item types are statistically significant at considerably beyond the .0001 level. The mean c for antonyms is higher than that for the three other verbal item types. Although the interaction is not statistically significant at any commonly accepted level, it is interesting to note that while for analogy items the mean c was .02 higher under formula-scoring instructions than under right-scoring instructions, for reading comprehension items the mean c was .01 lower under formula-scoring instructions.

DISCUSSION

It has often been noted that for most items c is less than 1/A, the probability that would be expected if a group of examinees guessed at random. Indeed, for three of the four verbal item types the mean c was 15 to 20 percent lower than the .20 that would have occured from random guessing (that is, the mean c was .17 or .16). For the antonym item type, however, the mean c was equal to 1/A. As is not unusual for an exploratory study, more questions were created than were answered.

1. It has been hypothesized (Lord, 1980), the finding that c tends to be less than 1/A indicates that many very low ability examinees do not guess at random, and tend to be mislead by plausible distractors. Does the finding that this is not affected by scoring instructions indicate that this is

a function of the same major dimension underlying test scores, or might there still be some other dimension(s), perhaps personality traits, that partially explain this phenomenon?

2. Do very low ability examinees guess at random for antonym items, but are they mislead into picking plausible distractors more frequently than would be accounted for by chance for the other three item types? It has been hypothesized (Petersen, personal communication) that if an examinee does not recognize the stem word, he or she will not be able to make use of either partial information or misinformation that exist in the distractors. For analogies, sentence completion, and reading comprehension items, however, numerous pieces of information are available in both the stem and the distractors.

3. What is it about antonym items that makes guessing behavior for them so different than for sentence completion items, even though for the GRE population, scores for the two item types correlate almost perfectly when corrected for attenuation (for example, for edition RS-H the uncorrected correlation between raw scores on antonyms and sentence completion was .71 and the correlation corrected for attenuation was .98)?

CONCLUSIONS AND RECOMMENDATIONS

Clearly more research is neccesary. A stronger research design would be useful, especially since analyses of c-parameters are essentially based on only a small portion of one's samples. Using the identical items for the two scoring conditions would have provided a more powerful analysis. Using the more recent version of LOGIST for both parts of the analysis would also have strengthened this research. This research has demonstrated, however, that there is a consistent mean difference between c-parameter estimates for different verbal item types. Thus, there is lilely to be advantage to an estimatation program that takes this into account. A Bayesian procedure could allow different priors for c's for different item types. Maximum likelihood approches, like LOGIST, could have separate common c estimates for those items for which insufficient data for individual c-estimates exist.

REFERENCES

Birnbaum, A. (1968). Some latent trait models. In Lord, F. & Novick, M. Statistical theories of mental test scores. Addison-Wesley: Reading, MA.

Box, G. (1954). Some theorems on quadratic forms applied to the study of analysis of variance problems: I Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 25, 290-302.

ETS (1984). Graduate Record Examinations Information Bulletin. Educational Testing Service: Princeton, NJ.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Erlbaum: Hillsdale, NJ.

Petersen, N. S. (1985) Personal communication.

Winer, B. J. (1971) Statistical principles in experimental design. McGraw-Hill: New York.

Wingersky, M. S. (1983) LOGIST: A program for computing maximum likelihood procedures for logistic test models. In Hambleton, R. K. (Ed.). Applications of item response theory. Educational Research Institute of British Columbia: Vancouver, British Columbia.

Table 1

Test Edition	Admin. Date	Number of Items				Mean Delta	Mean Score	Sample Size
		Anal.	Ant.	S.C.	R.C.			
FS-A	12/79	18	20	17	25	11.8	498	4,574
FS-B	2/80	18	20	17	25	11.8	472	4,475
FS-C	4/80	18	20	17	25	11.8	472	4,835
FS-D	6/80	18	22	13	22	11.8	473	2,984
RS-E	10/81	17	20	13	22	12.0	495	4,408
RS-F	12/81	18	22	14	22	11.8	496	4,096
RS-G	2/82	18	22	14	22	11.9	482	3,746
RS-H	4/82	18	22	14	22	11.9	465	3,647
RS-I	10/82	18	22	14	22	11.8	500	4,331
RS-J	4/83	18	22	14	22	11.9	485	3,825

Table 2

Standard Deviations of c-Parameter Estimates

Scoring Method	Analogies	Antonyms	Sentence Completion	Reading Comprehension
Formula	.04	.06	.05	.04
Right	.08	.08	.09	.08

Table 3

Means of c-Parameter Estimates

Scoring Method	Analogies	Antonyms	Sentence Completion	Reading Comprehension	Marginal[1]
Formula	.17	.20	.17	.16	.18
Right	.15	.20	.16	.17	.17
Marginal	.16	.20	.17	.16	.17

[1]Note, marginals are based on unweighted cell means.

Table 4

Analysis of Variance

Source of Variation	df	SS	MS	F	p
Item Type	3	.3421	.1140	23.23	<.01
Scoring Instructions	1	.0051	.0051	1.17	.28
Interaction	3	.0263	.0088	1.78	.15
Error	759	3.7252	.0049		

ACKNOWLEDGEMENTS

Data for this study came from two earlier studies funded by the Graduate Record Examinations Board and Educational Testing Service. An earlier version of this paper was presented at the 1985 Meeting of the National Coucil on Measurement in Education. This article has benefited from the two discussants, Robert Brennan and Ross Traub, as well as the review of colleagues at Educational Testing Service. The opionions expressed herein, however, are solely those of the author. The programming assistance of Louann Benton is gratefully acknowledged.